Bayesian Multidimensional Scaling and Choice of Dimension [1]

Man-Suk Oh
Ewha Women's University, Korea

Adrian E. Raftery
University of Washington

Technical Report no. 379
Department of Statistics
University of Washington
Seattle, WA 98195.

August 2000

[1]Man-Suk Oh is Associate Professor of Statistics, Department of Statistics, Ewha Women's University, Seoul 120-750, Korea. Email: msoh@mm.ewha.ac.kr. Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322. Email: raftery@stat.washington.edu; Web: www.stat.washington.edu/raftery. Oh's research was supported by a Research Fund provided by the Korean Research Foundation, Support for faculty Research Abroad. Raftery's research was supported by ONR Grant no. N00014-96-1-1092. The authors are very grateful to Chris Fraley for helpful comments and discussions. They are also grateful to Kate Stovel for providing the Lloyds Bank data and for helpful discussions about them.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2000		**00-08-2000 to 00-08-2000**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Bayesian Multidimensional Scaling and Choice of Dimension	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**31**	

Abstract

Multidimensional scaling is widely used to handle data which consist of dissimilarity measures between pairs of objects or people. We deal with two major problems in metric multidimensional scaling — configuration of objects and determination of the dimension of object configuration — within a Bayesian framework. A Markov chain Monte Carlo algorithm is proposed for object configuration, along with a simple Bayesian criterion for choosing their effective dimension, called MDSIC. Simulation results are presented, as well as examples on real data. Our method provides better results than classical multidimensional scaling for object configuration, and MDSIC seems to work well for dimension choice in the examples considered.

Key Words: Clustering, Dimensionality, Dissimilarity, Markov chain Monte Carlo, Metric scaling, Model selection.

Contents

1 Introduction **1**

2 Classical Multidimensional Scaling **3**

3 Bayesian Multidimensional Scaling **5**
3.1 Model and Prior 5
3.2 Markov Chain Monte Carlo 6
3.3 Posterior Inference 8

4 Dimensionality **9**

5 Examples **12**
5.1 A Simulation 12
5.2 Airline Distances between Cities 14
5.3 Careers of Lloyds Bank Employees, 1905–1950 15

6 Summary and Discussion **21**

List of Tables

1 Analysis of the simulation data in Example 1 13
2 Analysis of the City Data 15
3 Analysis of the LLoyd Bank data 20

List of Figures

1 Airline distances between cities 14
2 MCMC time sequence plots for the City data 16
3 Observed and estimated distances for the Airline distance data 17
4 Estimated locations of the cities from BMDS. 18
5 Fitted and observed dissimilarities for the LLoyd bank data 20
6 Pairwise scatter plots of the estimated object configuration from BMDS for the LLoyd Bank data. 22
7 Lloyds bank Data: First two BMDS dimensions 23

1 Introduction

Multidimensional scaling (MDS) is concerned with data that are given as dissimilarity measures between pairs of objects or individuals. Its goal is to represent the objects or individuals by points in a (usually) Euclidean space.

MDS has its roots in psychology, specifically psychophysics, being based on the analogy between the psychological concept of similarity, and the geometrical concept of distance. Two individuals are viewed as similar if they tend to have similar responses to the same stimuli. Subsequently, it has been widely used in other social and behavioral sciences. Recently, interest in MDS has increased further, due to its usefulness in some currently rapidly developed subjects, such as genomics (Tibshirani et al. 1999), and information retrieval for the Web and other document databases (Schutze and Silverstein, 1997).

One of the main applications of MDS is visualization, where the user wants to represent a complex set of dissimilarities in a form that is easier to see. One reason for this is to see if visually apparent clusters are present in the data. Another application is exploration, where the user wants to understand what the main dimensions underlying the dissimilarities are. For example, the objects in MDS might be political candidates, and the data might consist of subjective similarity judgements. MDS might help to suggest which political positions or characteristics are important in forming similarity judgements (e.g. position on Social Security, age, tendency to tell jokes). A third application is hypothesis testing. Monogaphs on MDS include Davison (1983), Young (1987), Cox and Cox (1994), and Borg and Groenen (1997).

An important issue in MDS is configuration of objects, i.e., estimation of values for attributes of objects. A commonly used MDS method for pairwise dissimilarity data was developed by Torgerson (1952, 1958). Object configurations are easy to compute with this method, now called *classical MDS* (CMDS). It gives complete recovery (up to location shift) of object configurations when the given dissimilarites are exactly equal to the Euclidean distances and when the dimension is correctly specified. In many practical situations, however, there are measurement errors in the observed dissimilarities and no clear notion of dimension. Maximum likelihood MDS methods have been developed for handling measurement errors — see, for instance, Ramsay (1982, 1991), Takane (1982), Takane and Carroll (1981), MacKay (1989), MacKay and Zinnes (1991), Groenen (1993), and Groenen, Mathar, and Heisser (1995). However, justification of maximum likelihood relies on asymptotic theory and computation requires nonlinear optimization. The number of parameters to be optimized over typically grows at a faster than linear rate relative to the dimension, so that the asymptotic theory may not apply in high dimensions, as pointed out by Cox (1982). Moreover,

the likelihood surface will tend to have many more local minima when there are more dimensions, and finding a good initial estimate will be correspondingly more difficult.

Another important issue in MDS is dimensionality, i.e., the number of significant attributes. Despite its importance in many applications, there is no definitive method for determining dimension for dissimilarity data. The most commonly used method is to search for an *elbow*, that is a point where a measure of fit or a measure of contribution to the dissimilarity levels off, in a plot of the measure versus dimension (Spence and Graef, 1974; Davison, 1983; Borg and Groenen, 1997). However, it is often difficult to find an elbow — especially when there are significant errors — and visual inspection of a plot may be misleading since its appearance often depends on the relative scale of the axes.

In this paper, we deal with these two important issues in MDS within a Bayesian framework. We use a Euclidean distance model and assume a Gaussian measurement error in the observed dissimilarity. Under the model, we propose a simple Markov chain Monte Carlo (MCMC) algorithm to obtain a Bayesian solution for the object configuration. We found that the proposed method, which we call *Bayesian MDS* (BMDS), provided a much better fit to the data than CMDS in all of the examples we tested. Moreover, the improvement in performance of the proposed BMDS scheme relative to CMDS was more pronounced when there were significant measurement errors in the data or when the Euclidean model assumption was violated.

Based on the BMDS estimate of object configuration over a range of dimensions, we propose a simple Bayesian criterion to choose an appropriate dimension. This criterion, called MDSIC, is based on the Bayes factor, or ratio of integrated likelihoods, for the BMDS estimated configuration under one dimension versus a different dimension. In simulated data, we found that the criterion works well for Euclidean models with measurement error that is not too large. In real examples, we found that the criterion gave satisfactory results. We also give an example of cluster analysis on real dissimilarity data in which the BMDS estimates of object configuration are used in conjunction with model-based clustering (Banfield and Raftery, 1993; Fraley and Raftery, 1998).

In our approach, observed dissimilarities are modeled as equal to Euclidean distances plus measurement error. In this sense, what we do here can be viewed as a Bayesian analysis of metric MDS, and here being Bayesian seems to confer the benefits of yielding good estimated configurations, and providing a formal way of choosing the dimension. A great deal of MDS research, however, has focused on *nonmetric* MDS, in which the relationship between dissimilarity and underlying distance is modeled as nonlinear. One could use the basic ideas here to do Bayesian nonmetric MDS, and we suggest some ways of doing this in Section 6.

The rest of the paper is organized as follows. Classical MDS (Torgerson 1952, 1958) is briefly

reviewed in Section 2. Bayesian MDS is described in Section 3: Section 3.1 presents the model and the prior, Section 3.2 presents an MCMC algorithm, and Section 3.3 describes the estimation of object configuration from the MCMC output. Based on the BMDS output, a simple Bayesian dimension selection criterion, MDSIC, is described in Section 4. Some simulated and real examples are given in Section 5. We conclude with a summary and discussion in Section 6.

2 Classical Multidimensional Scaling

In this section, we briefly review classical MDS. Let δ_{ij} denote the dissimilarity measure between objects i and j, which are functionally related to p unobserved attributes of the objects. Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denote an unobserved vector representing the values of the attributes possesed by object i.

Torgerson (1952, 1958) developed a technique for multidimensional scaling, now called classical MDS. Assume that the dissimilarity measure, δ_{ij}, is the distance between objects i and j in a p-dimensional Euclidean space, i.e.,

$$\delta_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{1}$$

where x_{ik} is the k-th element of $\mathbf{x}_i$. The elements x_{ik} are unknown, and the goal of MDS is to recover them from the dissimilarity data. Because of the non-identifiability of the solution under location shift, the center of the object points is placed at the origin, so that $\sum_{i=1}^{n} x_{ij} = 0$ for $j = 1, ..., p$, where n is the total number of objects.

Construct a double-centered matrix A with elements a_{ij} defined by

$$a_{ij} = -\frac{1}{2}(\delta_{ij}^2 - \delta_{i\cdot}^2 - \delta_{\cdot j}^2 + \delta_{\cdot\cdot}^2),$$

where

$$\delta_{i\cdot}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2, \quad \delta_{\cdot j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2, \quad \delta_{\cdot\cdot}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2.$$

It was shown by Togerson (1952, 1958) that

$$a_{ij} = \sum_{k=1}^{p} x_{ik} \cdot x_{jk}, \text{ for all } i, j, \text{ i.e., } A = \mathbf{X}\mathbf{X}', \tag{2}$$

where $\mathbf{X}$ is the $n \times p$ matrix of object coordinates. The coordinates of $\mathbf{X}$ can be recovered from the spectral decomposition of the matrix A in (2). If the observed dissimilarities, d_{ij}, satisfy the Euclidean distance assumption and there is no measurement error, then the Euclidean distances

computed from the matrix $\mathbf{X}$ satisfying (2) will be exactly equal to the given dissimilarities. However, when the model assumption is violated or when there are significant measurement errors in the data, CMDS estimates of object configuration may not be very useful.

It should be noted that the matrix $\mathbf{X}$ satisfying (2) is not unique, because Euclidean distance is invariant under translation, rotation, and reflection about the origin. However, there is a unique CMDS solution having zero mean, diagonal covariance, and some fixed coordinate signs.

At present, there is no definitive method for choosing the effective dimension of $\mathbf{x}_i$, the number of object attributes that contribute significantly to the dissimilarities. A common way of assessing dimension is to look at the eigenvalues of the scalar product matrix A. The k-th eigenvalue is a measure of contribution of the k-th coordinate of $\mathbf{X}$ to squared distances. Small eigenvalues (relative to the largest eigenvalue) imply that the corresponding coordinates make little contribution to the squared distances and hence only the first p coordinates of $\mathbf{X}$ corresponding to the first p significantly large eigenvalues suffice to represent the objects. To determine significantly large eigenvalues, one may draw a plot of the ordered eigenvalues versus dimension and look for a dimension at which the sequence of eigenvalues levels off. If each δ_{ij} is equal to a p-dimensional Euclidean distance between objects i and j as given in (1), then the plot should level off precisely at dimension $(p+1)$.

A measure of fit, called *stress*, is also commonly used to determine the dimensionality. Several definitions of stress have been proposed; the one we use here, and perhaps the mostly commonly used one, is

$$STRESS = \sqrt{\frac{\sum_{i \neq j}(d_{ij} - \hat{\delta}_{ij})^2}{\sum_{i \neq j} d_{ij}^2}},$$

where $\hat{\delta}_{ij}$ is the Euclidean distance obtained from the estimated object configuration (Kruskal 1964). A plot of STRESS versus dimension will level off at the true dimension p, if $d_{ij} = \hat{\delta}_{ij}$ and $\hat{\delta}_{ij}$ is given by (1).

Both methods rely on detecting an *elbow* in a sequence of values, that is, a point where the plot levels off. However, in real data that do not conform exactly to the model or in which there is a significant amount of measurement or sampling error, an elbow may be difficult to discern. In addition, visual detection of an elbow in a plot can be misleading since the outcome may depend on the scale of the axes.

3 Bayesian Multidimensional Scaling

3.1 Model and Prior

Dissimilarity data can be obtained in various forms. However, since Euclidean distance is easy to handle and is relatively insensitive to the choice of dimension compared to other distance measures, it tends to be used in cases in which the dimension is unknown unless there is strong theoretical evidence for preferring a non-Euclidean distance (Davison, 1983). Thus, for Bayesian MDS we model the true dissimilarity measure δ_{ij} as the distance between objects i and j in a Euclidean space, as given in (1).

In practical situations there are often measurement errors in observations. In addition, dissimilarity measures are typically given as positive values. We therefore assume that the observed dissimilarity measure d_{ij} comprising the data is equal to the true measure δ_{ij} plus a Gaussian error, with the restriction that the observed dissimilarity measure is always positive. In other words, given δ_{ij}, the observed dissimilarity measure d_{ij} is assumed to follow the truncated normal distribution

$$d_{ij} \sim N(\delta_{ij}, \sigma^2)\ I(d_{ij} > 0), \quad i \neq j, i, j = 1, ..., n, \tag{3}$$

where $\delta_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$, and the x_{ik} are unobserved. From this, the likelihood function of the unknown parameters $\mathbf{X} = \{\mathbf{x}_i\}$ and σ^2 is

$$l(\mathbf{X}, \sigma^2) \propto (\sigma^2)^{-m/2}\ \exp\left[-\frac{1}{2\sigma^2}SSR - \sum_{i>j}\log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right], \tag{4}$$

where $m = n(n-1)/2$ is the number of dissimilarities, $SSR = \sum_{i>j}(d_{ij} - \delta_{ij})^2$ is the sum of squared residuals, and $\Phi(\cdot)$ is the standard normal cdf.

For Bayesian analysis of the model, we need to specify priors for $\mathbf{X} = \{\mathbf{x}_i\}$, and σ^2. For the prior distribution of $\mathbf{x}_i$, we use a multivariate normal distribution with mean 0 and a diagonal covariance matrix Λ, i.e.,

$$\mathbf{x}_i \sim N(0, \Lambda),$$

independently for $i = 1, .., n$. For the prior of the error variance σ^2, we use a conjugate prior

$$\sigma^2 \sim IG(a, b),$$

the inverse Gamma distribution with mode $b/(a+1)$. For a hyper-prior for the elements of $\Lambda = \text{Diag}(\lambda_1, ..., \lambda_p)$, given dimension p, we also assume a conjugate prior,

$$\lambda_j \sim IG(\alpha, \beta_j),$$

independently for $j = 1, .., p$. We assume prior independence among $\mathbf{X}$, Λ, and σ^2, i.e., $\pi(\mathbf{X}, \sigma^2, \Lambda) = \pi(\mathbf{X})\ \pi(\sigma^2)\ \pi(\Lambda)$, where $\pi(\mathbf{X})$, $\pi(\sigma^2)$, and $\pi(\Lambda)$ are the priors given above.

When there is little prior information, one may use either the results from a preliminary run or the results from CMDS for parameter specification in the priors. For instance, one may choose a small a for a vague prior of σ^2 and choose b so that the prior mean matches with the estimated mean of σ^2 from CMDS. Similarly, for the hyper-prior of λ_j, one may choose a small α and choose β_j so that the prior mean of λ_j matches with the j-th diagonal element of the covariance matrix $S_x = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i$ obtained from CMDS. As noted above, the CMDS solution can be transformed so as to have zero mean and diagonal covariance.

Such a prior is mildly data-dependent, and it might be argued that this violates the definition of a prior distribution. However, we view this prior as an approximation to the elicited data-independent prior of an analyst who knows a little, but not much, about the problem at hand. Because this prior is diffuse relative to the likelihood, the estimation results are unlikely to be highly sensitive to its precise specification.

3.2 Markov Chain Monte Carlo

From the likelihood and the prior, the posterior density function of the unknown parameters $(\mathbf{X}, \sigma^2, \Lambda)$ is

$$\begin{aligned}\pi(\mathbf{X}, \sigma^2, \Lambda|D) \quad &\propto \quad (\sigma^2)^{-(m/2+a+1)} \\ &\times \prod_{j=1}^{p} \lambda_j^{-n/2}\ \exp\left[-\frac{1}{2\sigma^2}SSR - \sum_{i>j} \log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right) - \frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i - \frac{b}{\sigma^2} - \sum_{j=1}^{p}\frac{\beta_j}{\lambda_j}\right] ,\end{aligned} \tag{5}$$

where D is the matrix of observed dissimilarities. Because of the complicated form of the posterior density function (5), numerical integration is required to obtain a Bayes estimate of the parameters. In particular, the posterior is a complicated function of the $\mathbf{x}_i$'s, which in most cases are of high dimension. We therefore use a Markov chain Monte Carlo (MCMC) algorithm (e.g. Gilks et al. 1996) to simulate from the posterior distribution (5). Our algorithm proceeds by iteratively generating new values for each object configuration $\mathbf{x}_i$, the error variance σ^2, and the hyperparameter Λ, given the current values of the other unknowns.

We first suggest initialization strategies for the unknown parameters which are needed for the MCMC algorithm. For initialization of $\mathbf{x}_i$, one may use the output, $\mathbf{x}_i^{(0)}$, of $\mathbf{x}_i$ from CMDS since it is easy to obtain. The resulting $\mathbf{X}$ can then be centered at the origin and then transformed using the spectral decomposition so as to have a diagonal covariance matrix, thus conforming to the prior. From the adjusted $\mathbf{x}_i^{(0)}$'s, one can compute the sum of squared residuals $SSR^{(0)}$ and an estimate $\sigma^{(0)2} = SSR^{(0)}/m$ which can be used as an initial value of σ^2 in the algorithm. In

addition, diagonal elements of the adjusted sample covariance matrix of $\mathbf{X}$ can be used as initial values for the λ_j's.

We now describe the details of sample generation in the MCMC algorithm. At each iteration, we simulate a new value of λ_j from its conditional posterior distribution given the other unknowns. From (5), the full conditional posterior distribution of λ_j is the inverse Gamma distribution $IG(\alpha + n/2, \beta_j + s_j/2)$, where s_j/n is the sample variance of the j-th coordinates of $\mathbf{x}_i$'s. We use a random walk Metropolis-Hastings step (Hastings, 1970) to generate $\mathbf{x}_i$ and σ^2 in each iteration of the MCMC algorithm.

Generation of $\mathbf{x}_i$

A normal proposal density is used in the random walk Metropolis-Hastings algorithm for generation of $\mathbf{x}_i$. To choose the variance of the normal proposal density, we note that the full conditional posterior density of $\mathbf{x}_i$ is

$$\pi(\mathbf{x}_i|\text{else}) \propto \exp\left[-\frac{1}{2}(Q_1 + Q_2) - \sum_{j\neq i, j=1}^{n} \log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right],$$

where

$$Q_1 = \frac{1}{\sigma^2}\sum_{j\neq i, j=1}^{n}(\delta_{ij} - d_{ij})^2; \qquad Q_2 = \mathbf{x}_i'\Lambda^{-1}\mathbf{x}_i.$$

Since

$$\begin{aligned}\frac{1}{\sigma^2}(\delta_{ij} - d_{ij})^2 &= \frac{1}{\sigma^2}\left(\delta_{ij}^2 - 2d_{ij}\delta_{ij} + d_{ij}^2\right) \\ &= \frac{1}{\sigma^2}\left[(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) - 2d_{ij}\sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)} + d_{ij}^2\right]\end{aligned}$$

is a quadratic function of $\mathbf{x}_i$ with leading coefficient equal to $1/\sigma^2$ and Q_1 has $n-1$ of this kind while Q_2 has only one quadratic term with coefficient Λ, Q_1 would dominate the full conditional posterior density function of $\mathbf{x}_i$ unless there is strong prior information. Thus, we may consider Q_1 only and approximate the full conditional variance of $\mathbf{x}_i$ by $\sigma^2/(n-1)$ and choose the variance of the normal proposal density to be a constant multiple of $\sigma^2/(n-1)$. With this proposal density, the random walk M-H algorithm can be summarized as:

- Generate $\mathbf{x}_i^N$ from $N\left(\mathbf{x}_i^C, k * \frac{\sigma^2}{(n-1)}\right)$, where $\mathbf{x}_i^C$ is the current sample of $\mathbf{x}_i$ and k is a constant.
- Replace $\mathbf{x}_i^C$ by $\mathbf{x}_i^N$ with probability $\min\{1, \pi(\mathbf{x}_i^N|\text{else})/\pi(\mathbf{x}_i^C|\text{else})\}$.

Generation of σ^2

From a preliminary numerical study, we found that

$$\pi(\sigma^2|\text{else}) \propto (\sigma^2)^{-(m/2+a+1)} \exp\left[-\frac{1}{\sigma^2}(SSR/2+b) - \sum_{i>j} \log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]$$

is well approximated by the density function of $IG(m/2+a, SSR/2+b)$, up to a constant of proportionality. When the number of dissimilarities, $m = n(n-1)/2$, is large, which is often the case since m is a quadratic function of n, the inverse Gamma density function is well approximated by a normal density. Thus, we propose a random walk M-H algorithm with a normal proposal density with variance proportional to the variance γ^2 of $IG(m/2+a, SSR/2+b)$ distribution, namely:

- Generate σ^{2N} from $N(\sigma^{2C}, k*\gamma^2)$, where σ^{2C} is the current sample of σ^2.
- Replace σ^{2C} by σ^{2N} with probability $\min\{1, \pi(\sigma^{2N}|\text{else})/\pi(\sigma^{2C}|\text{else})\}$.

3.3 Posterior Inference

Iterative generation of $\{\mathbf{x}_i, i = 1, .., n\}$, σ^2, and $\lambda_j, j = 1, .., p$, for a sufficiently long time provides a sample from the posterior distribution of the unknown parameters, and Bayes estimation of the parameters can be obtained from the samples. However, because the model assumes a Euclidean distance for the dissimilarity measure δ_{ij}, the posterior samples of $\{\mathbf{x}_i\}$ would be invariant under translation, rotation and reflection about the origin, as in classical MDS, unless there is strong prior information to the contrary. We can retrieve only the relative locations of the objects from the data, and not their absolute locations. Hence the convergence of δ_{ij} rather than $\mathbf{X}$ needs to be checked to verify the convergence of MCMC. The near lack of identifiability in $\mathbf{X}$ also suggests the use of sample averages as Bayes estimates to be inadvisable, since the MCMC samples of $\mathbf{X}$ may be unstable even when the distances δ_{ij} are stable. Thus, we take an approximate posterior mode of $\mathbf{X}$ as a Bayes estimate of $\mathbf{X}$, i.e., the BMDS solution of the object configuration. The posterior mode provides relative positions of $\mathbf{x}_i$'s corresponding to the maximum posterior density. A meaningful absolute position of $\mathbf{X}$ may be obtained from an appropriate transformation, if desired.

To obtain the posterior mode of $\mathbf{X}$, one may compute the posterior kernel for each MCMC sample. However, this can be time consuming, since the posterior is complicated. However, we observed that the likelihood dominates the prior and in the likelihood (4) the term involving SSR is dominant, so that the posterior mode of $\mathbf{X}$ can be approximated by the value of $\mathbf{X}$ which minimizes the sum of squared residuals SSR.

Since the center and direction of $\mathbf{X}$ can be arbitrary, we post-process the MCMC sample of $\mathbf{X}$ at each iteration using the transformation

$$\mathbf{x}_i = D_x'(\mathbf{x}_i - \bar{\mathbf{x}}),$$

where $\bar{\mathbf{x}}$ is the average of $\mathbf{x}_i$'s and D_x is the matrix whose columns are the eigenvectors of the sample covariance matrix $S_x = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})$ of $\mathbf{x}_i$'s. This transformation does not solve the non-indentifiability problem but the new $\mathbf{x}_i$'s have mean 0 and a diagonal covariance matrix to correspond to the prior specification.

4 Dimensionality

BMDS as described in the previous section gives object configurations in a given dimensional Euclidean space. In most cases the dimension of the objects (the number of significant attributes) is unknown. In this section, we propose a simple Bayesian dimension selection criterion based on the BMDS object configurations.

Consider the dimension p as an unknown variable and assume equal priors for all values of p. Then the posterior is given as

$$\begin{aligned}
&\pi(\mathbf{X}, \sigma^2, \Lambda, p|D) \propto l(\mathbf{X}, \sigma^2, p|D)\pi(\mathbf{X}|\Lambda, p)\pi(\sigma^2)\pi(\Lambda|p) \\
&= (2\pi)^{-m/2}\sigma^{-m} exp[-\frac{1}{2\sigma^2}SSR - \sum_{i>j} log\Phi(\delta_{ij}/\sigma)] \cdot (2\pi)^{-np/2}\prod_{j=1}^{p}\lambda_j^{-n/2} exp[-\sum_{j=1}^{p}\frac{1}{2\lambda_j}s_j] \\
&\quad \times \Gamma(a)^{-1}b^a(\sigma^2)^{-(a+1)}exp[-b/\sigma^2] \cdot \Gamma(\alpha)^{-p}\prod_{j=1}^{p}\beta_j^{\alpha}\lambda_j^{-(\alpha+1)}exp[-\beta_j/\lambda_j] \\
&= A \cdot h(\sigma^2, \mathbf{X}) \cdot g(\Lambda, \mathbf{X}),
\end{aligned}$$

where

$$s_j = \sum_{i=1}^{n} x_{ij}^2, \tag{6}$$

$$A = (2\pi)^{-(m+np)/2}\ \Gamma(a)^{-1}b^a\ \Gamma(\alpha)^{-p}\ \prod_{j=1}^{p}\beta_j^{\alpha}, \tag{7}$$

$$h(\sigma^2, \mathbf{X}) = (\sigma^2)^{-(m/2+a+1)}\ \exp\left[-(SSR/2 + b)/\sigma^2\right], \tag{8}$$

$$g(\Lambda, \mathbf{X}) = \prod_{j=1}^{p}\lambda_j^{-(n/2+\alpha+1)}\ \exp\left[-(s_j/2 + \beta_j)/\lambda_j\right]. \tag{9}$$

Note that, because of the post-processing described in Section 3.3, the $\mathbf{x}_i$'s have mean 0 and a diagonal covariance matrix.

We adopt a Bayesian approach to choosing the dimension. We view the overall task to be that of choosing the best configuration, and hence we view the choice between dimension p and dimension p' as being between the estimated configuration with dimension p and the estimated configuration with dimension p'. Thus, we consider the marginal posterior, $\pi(\mathbf{X}, p|D)$, of $(\mathbf{X}, p)$ with $\mathbf{X}$ equal to the BMDS solution and choose the value of p which gives the largest value of $\pi(\mathbf{X}, p|D)$. When choosing between dimension p and dimension p', this is based on the posterior odds for the estimated configuration of dimension p as against the estimated configuration of dimension p'.

Now, to compute the criterion, note that

$$\begin{aligned} \pi(\mathbf{X}, p|D) &\propto \int l(\mathbf{X}, \sigma^2, p|D)\pi(\sigma^2)d\sigma^2 \cdot \int \pi(\mathbf{X}, \Lambda, p)d\Lambda \\ &\approx l(\mathbf{X}, p|D)\pi(\mathbf{X}, p), \end{aligned}$$

where $l(\mathbf{X}, p|D)$is the marginal likelihood of $(\mathbf{X}, p)$ and $\pi(\mathbf{X}, p)$ is the marginal prior of $(\mathbf{X}, p)$. The marginalised likelihood term would increase as p increases. However, the marginal prior term decreases as p increases, since we are using a diffuse (but proper) prior, and so this term penalizes more complex models. The approach has the simplicity of a maximum likelihood method, as well as the advantage of a Bayesian method in penalizing more complex models.

Integrating the function $g(\Lambda, \mathbf{X})$ given in (9) with respect to Λ gives

$$\int g(\Lambda, \mathbf{X})d\Lambda = \Gamma^p(n/2 + \alpha) \prod_{j=1}^{p} (s_j/2 + \beta_j)^{-(n/2+\alpha)}. \tag{10}$$

The integral of the function $h(\sigma^2, \mathbf{X})$ given in (8) with respect to σ^2 is approximately equal to

$$\int h(\sigma^2, \mathbf{X})d\sigma^2 \approx (2\pi)^{1/2}(m/2)^{-1/2}(SSR/m)^{-m/2+1} \exp[-m/2]. \tag{11}$$

This formula is justified in the Appendix. From these,

$$\begin{aligned} \pi(\mathbf{X}, p|D) &\propto A \cdot \int h(\sigma^2)d\sigma^2 \cdot \int g(\Lambda, \mathbf{X})d\Lambda \\ &= A^* \cdot (SSR/m)^{-m/2+1} \cdot \prod_{j=1}^{p} (s_j/2 + \beta_j)^{-(n/2+\alpha)}, \end{aligned} \tag{12}$$

where

$$A^* = A \cdot (2\pi)^{1/2}\ \Gamma^p(n/2 + \alpha)\ (m/2)^{-1/2}\ \exp[-m/2].$$

To clarify the dependence of $\mathbf{X}$ on p, let $\mathbf{X}^{(p)}$ denote the BMDS solution of $\mathbf{X}$ when the dimension is p. There is a difficulty in directly comparing $(\mathbf{X}^{(p)}, p)$ and $(\mathbf{X}^{(p+1)}, p+1)$. The marginal posterior $\pi(\mathbf{X}, p|D)$ is dependent on the scale of $\mathbf{X}$, because it includes the term $\prod_{j=1}^{p}(s_j/2 + \beta_j)^{-(n/2+\alpha)}$. Note that s_j/n is the sample variance of the j-th coordinate of $\mathbf{X}$. However, without improvement

in the fit, the scale of $\mathbf{X}$ may change with the dimension p. Given the same Euclidean distances, the coordinates of $\mathbf{X}$ would get closer to the origin as p increases, unless all the extra coordinates are equal to 0. For instance, the Euclidean distance between -1 and 1 in one-dimensional space is equal to the Euclidean distance between $(1/\sqrt{2}, 1/\sqrt{2})$ and $(-1/\sqrt{2}, -1/\sqrt{2})$ in two-dimensional space, and hence the variance in each coordinate is smaller in two-dimensional space. This would give a smaller s_j and hence a larger $\pi(\mathbf{X}, p|D)$ in a higher dimension, even though there is no change in the distance and the fit.

To circumvent this scale dependency, a dimension selection criterion should compare $\mathbf{X}$'s in the same dimension. For this, let $\mathbf{X}^{*(p+1)} = (\mathbf{X}^{(p)} : \mathbf{0})$ in $(p+1)$-dimensional space, which has the first p coordinates equal to $\mathbf{X}^{(p)}$ and the last coordinates all equal to 0. Then $\mathbf{X}^{*(p+1)}$ provides the same Euclidean distances and the same fit as $\mathbf{X}^{(p)}$ and may be considered an implantation of $\mathbf{X}^{(p)}$ in $(p+1)$-dimensional space. Ideally, if p is the correct dimension, then the optimal solution $\mathbf{X}^{(p+1)}$ in $(p+1)$-dimensional space would be $\mathbf{X}^{*(p+1)}$. Thus, we compare $\mathbf{X}^{*(p+1)}$ and $\mathbf{X}^{(p+1)}$ and choose p to be the dimension if $\mathbf{X}^{*(p+1)}$ has a larger marginal posterior density than $\mathbf{X}^{(p+1)}$.

From (12), the ratio of the marginal posteriors of $\mathbf{X}^{*(p+1)}$ and $\mathbf{X}^{(p+1)}$ is

$$\begin{aligned} R_p &\equiv \frac{\pi(\mathbf{X}^{(p+1)}, p+1|D)}{\pi(\mathbf{X}^{*(p+1)}, p+1|D)} \\ &= \left(\frac{SSR_{p+1}}{SSR^*_{p+1}}\right)^{-m/2+1} \left(\prod_{j=1}^{p+1} \frac{s_j/2+\beta_j}{s^*_j/2+\beta_j}\right)^{-(n/2+\alpha)} \\ &= \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-m/2+1} \left(\prod_{j=1}^{p} \frac{s_j^{(p+1)}/2+\beta_j}{s_j^{(p)}/2+\beta_j}\right)^{-(n/2+\alpha)} \left(\frac{s_{p+1}^{(p+1)}/2+\beta_j}{\beta_j}\right)^{-(n/2+\alpha)}, \end{aligned}$$

where $s_j^{(p)}$ is s_j given in (6), computed from $\mathbf{X}^{(p)}$. Clearly, the ratio R_p depends on the choice of the hyper-parameters α and β_j of Λ.

When there is no strong prior information, a reasonable choice for α, β_j in $(p+1)$-dimensional space might be $\alpha = \frac{1}{2}$ and $\beta_j = \frac{1}{2}s_j^{(p+1)}/n$ so that the prior information roughly corresponds to the information from one observation. This is close to the unit information prior, which was observed by Kass and Wasserman (1995) to correspond to the BIC approximation to the Bayes factor (Schwarz 1978), and by Raftery (1995) to correspond to a similar approximation to the integrated likelihood. Raftery (1999) argued that this is a reasonable proper prior for approximating the situation where the amount of prior information is small.

This yields the ratio

$$R_p = \left(\frac{SSR_{p+1}}{SSR_p}\right)^{-m/2+1} \left(\prod_{j=1}^{p} \frac{r_j^{(p+1)}(n+1)}{(n+r_j^{(p+1)})}\right)^{-(n+1)/2} (n+1)^{-(n+1)/2},$$

where $r_j^{(p+1)} = s_j^{(p+1)}/s_j^{(p)}$. Taking minus twice the log of the ratio gives

$$\begin{aligned} LR_p &\equiv -2 \log R_p \\ &= (m-2)\log(SSR_{p+1}/SSR_p) \qquad (13) \\ &\quad + \left\{ (n+1)\sum_{j=1}^{p} \log\left[\frac{r_j^{(p+1)}(n+1)}{(n+r_j^{(p+1)})}\right] + (n+1)\ \log(n+1) \right\}. \qquad (14) \end{aligned}$$

Note that the term (13) in LR_p is roughly the log likelihood ratio, and would be negative since higher dimension results in a smaller SSR. The term (14) plays the role of penalty on the increase of dimension by one and would be positive if $r_j^{(p+1)} \leq 1$ and $\prod_{j=1}^{p} r_j^{(p+1)} > 1/(n+1)$. When there is no significant change in $\mathbf{X}$ between p- and $(p+1)$-dimensional spaces, then $r_j^{(p+1)} \approx 1$ and the penalty term is approximately $(n+1)\ \log(n+1)$.

A positive LR_p would prefer the dimension p to $(p+1)$ and a negative value would prefer the dimension $(p+1)$ to p and hence one can select the dimension where the value of LR_p turns positive. Alternatively, if we define $MDSIC$ as

$$\begin{aligned} MDSIC_1 &= (m-2) log\ SSR_1 \qquad (15) \\ MDSIC_p &= MDSIC_1 + \sum_{j=1}^{p-1} LR_j \qquad (16) \end{aligned}$$

then the optimal dimension is the one which achieves the minimum of $MDSIC_p$.

5 Examples

BMDS requires that prior parameters be specified. For all the examples given in this section, we chose 5 degrees of freedom a for the prior of σ^2 and chose b to match the prior mean of σ^2 with the estimate obtained from the CMDS. Note that a smaller a would not make much difference since $m = n(n-1)/2$ is large. For the hyper-prior of λ_j, we choose $\alpha = 1/2$ and $\beta_j = \frac{1}{2}s_j^{(0)}/n$, where $s_j^{(0)}/n$ is the estimated variance of the j-th coordinate of $\mathbf{X}$ obtained from the CMDS, which roughly corresponds to information from one observation as described in Section 4.

For the constant k in the Metropolis-Hastings algorithms for generating $\mathbf{x}_i$ and σ^2, we chose $k = 2.38^2$ for both $\mathbf{x}_i$ and σ^2 as suggested by Gelman et al. (1996). We found reasonably fast mixing in MCMC with this choice of k.

5.1 A Simulation

As an illustrative example, we generated 50 random samples of $\mathbf{x}_i$ from a 10-dimensional multivariate normal distribution with mean 0 and variance I, the identity matrix. We used the Euclidean

dim	CMDS STRESS	BMDS STRESS	LRT	Penalty	MDSIC
1	0.6622	0.4864	-1118.7	177.0	10673
2	0.4943	0.3078	-830.4	170.7	9731
3	0.3720	0.2192	-693.0	165.3	9071
4	0.2751	0.1651	-638.1	164.2	8544
5	0.2037	0.1272	-499.2	174.2	8070
6	0.1580	0.1037	-535.5	171.4	7745
7	0.1092	0.0833	-334.0	178.6	7381
8	0.0809	0.0727	-237.6	177.8	7225
9	0.0672	0.0660	-195.2	182.2	7165
10	0.0614	0.0609	-23.6	196.8	7152*
11	0.0658	0.0603	12.3	203.4	7326
12	0.0715	0.0606	-22.8	195.0	7541
13	0.0784	0.0601	2.5	232.7	7713
14	0.0855	0.0601	-29.8	170.5	7949

Table 1: Analysis of the simulation data in Example 1, $\mathbf{x}_i \sim N_{10}(0, I)$. The minimum MDSIC is marked with a star.

distances between pairs of $\mathbf{x}_i, \mathbf{x}_j$ as dissimilarities δ_{ij}. Given these δ_{ij}'s, we generated the observed distances d_{ij} from a normal distribution with mean δ_{ij} and standard deviation 0.3, truncated at 0. Thus, the data consist of a 50×50 symmetric matrix of dissimilarities computed from Euclidean distances with Gaussian errors.

Using the results from CMDS for initialization, BMDS as described in Section 3 was applied for various values of the dimension p. Time sequence plots from samples of δ_{ij}'s and σ^2 in MCMC converged quickly, similarly to Figure 2. We took samples from 10000 consecutive iterations after 3000 burn-in iterations from MCMC. With minimum SSR and $\mathbf{x}_i$ obtained from the BMDS, we applied MDSIC described in Section 4 to select the dimension of $\mathbf{x}_i$. The results are summarized in Table 1.

The first and the second columns show values of STRESS from CMDS and BMDS, respectively. The third and the fourth columns show the likelihood ratio term of (13) and the penalty term of (14), respectively. The last column shows the MDSIC given in (16). It can be observed that BMDS gives a better fit than CMDS, providing a smaller STRESS, especially when the dimension is incorrect. This is interesting because, for visualization purposes, dimension $p = 2$ is often chosen. In this case, the STRESS from CMDS for dimension 2 is 60% greater than for BMDS.

It is interesting to note that, for CMDS, STRESS increases after dimension 10 while for BMDS, STRESS stays roughly constant after dimension 10. Ideally, since the true dimension is 10, all

									NY	22	36	48	43	26	51	24	73	100	68
Azores	AZ									PY	54	33	59	33	31	37	93	88	84
Bagdad	39	BD									PS	57	7	56	72	50	57	105	61
Berlin	22	20	BN									RO	57	66	18	69	113	84	115
Bombay	59	20	39	BY									RE	63	74	57	57	101	61
BuenosAires	54	81	74	93	BS									SF	59	7	61	74	52
Cairo	33	8	18	27	73	CO									SO	64	117	71	107
Capetown	57	49	60	51	43	45	CN									SE	57	77	48
Chicago	32	64	44	81	56	61	85	CH									SI	49	11
Guam	89	63	71	48	104	71	88	74	GM									SY	48
Honolulu	73	84	73	80	76	88	115	43	38	HU									TO
Istanbul	29	10	11	30	76	8	52	55	69	81	IL								
Juneau	46	61	46	69	77	63	103	23	51	28	55	JU							
London	16	25	6	45	69	22	60	40	75	72	16	44	LN						
Manila	83	49	61	32	111	57	75	81	16	53	57	59	67	MA					
Melbourne	120	81	99	61	72	87	64	97	35	55	91	81	105	39	ME				
Mexico City	45	81	61	97	46	77	85	17	75	38	71	32	56	88	84	MY			
Montreal	24	58	37	75	56	54	79	8	77	49	48	26	33	82	104	23	ML		
Moscow	32	16	10	31	84	18	63	50	61	70	11	46	16	51	90	67	44	MW	
New Orleans	36	72	51	89	49	68	83	8	77	42	62	29	46	87	93	9	14	58	NS
New York	25	60	40	78	53	56	78	7	80	50	50	29	35	85	104	21	3	47	12
Panama City	38	78	59	97	33	71	70	23	90	53	68	45	53	103	90	15	25	67	16
Paris	16	24	5	44	69	20	58	41	76	75	14	47	2	67	104	57	34	16	48
RioDeJaneiro	43	69	62	83	12	61	38	53	116	83	64	76	57	113	82	48	51	72	48
Rome	21	18	7	38	69	13	52	48	76	80	9	53	9	65	99	64	41	15	55
SanFrancisco	50	75	57	84	64	75	103	19	58	24	67	15	54	70	79	19	25	59	19
Santiago	57	88	78	100	7	80	49	53	98	69	81	73	72	109	70	41	54	88	45
Seattle	46	68	51	77	69	68	102	17	57	27	61	9	48	67	82	23	23	52	21
Shanghai	72	44	51	31	122	52	81	70	19	49	49	49	57	12	50	80	70	42	77
Sydney	121	83	100	63	73	90	69	92	33	51	93	77	106	39	4	81	100	90	89
Tokyo	73	52	56	42	114	60	92	63	16	39	56	40	60	19	51	70	65	47	69

Figure 1: Airline distances between cities (100 miles) (The World Almanac, 1966).

coordinates except for the first 10 should be roughly zero and SSR should be about the same after dimension 10. This is precisely what ocurred in BMDS, indicating that BMDS handles the measurement errors well.

The third and the fourth columns show that the log likelihood ratios computed from the BMDS solution for $\mathbf{X}$ decrease monotonically as p increases upto 10, that there is no significant change after dimension 10, and that the penalties stay about the same for various p. Moreover, the MDSIC assumes a minimum at the correct dimension, namely 10.

5.2 Airline Distances between Cities

The World Almanac (1966, p.510) provided airline distances between 30 principal cities of the world, which are shown in Figure 1. Cities are located on the surface of the earth, a 3-dimensional sphere, and airplanes travel on the surface of the earth. Thus, airline distances are not exactly Euclidean distances and we may expect the dimension of $\mathbf{x}_i$ to be between 2 and 3.

The BMDS is applied to the data. Figure 2 shows time sequence plots of SSR and some δ_{ij}'s from 10000 MCMC iterations after a burn-in of 3000. The plots suggest that convergence has been achieved. Here we are interested in finding an approximate posterior mode of $\mathbf{X}$ rather than its full posterior distribution, so that convergence requirements are less stringent than if one seeks the full posterior distribution. The BMDS results are shown in Table 2.

Bayesian MDS yielded much smaller SSR than classical MDS in all cases. The estimated SSR

dim	CMDS STRESS	BMDS STRESS	LRT	Penalty	MDSIC
1	0.6782	0.3617	-704.2	95.7	5336
2	0.4682	0.1604	-548.5	91.4	4727
3	0.3811	0.0851	4.7	108.0	4270*
4	0.4006	0.0856	-2.4	88.9	4383
5	0.4139	0.0854	4.1	143.9	4469

Table 2: Analysis of the City Data. The minimum MDSIC is marked with a star.

dropped very quickly until dimension 3 and then increased slightly at dimension 4. MDSIC selected dimension 3. We observed that the last coordinates of $\mathbf{x}_i$ in dimension 4 are almost equal to 0, indicating strong evidence for dimension 3.

Figure 3 is a plot of the observed airline distances versus the estimated Euclidean distances. A perfect fit would yield a forty-five degree line as shown in Figure 3. The estimated Euclidean distances from BMDS are represented as red dots and and those from CMDS as green dots. One can see that BMDS provided points very close to the forty-five degree line except for the points corresponding to large distances. The fit gets worse as the distance gets larger, because when cities are farther apart, there is a greater discrepancy between airline distance and 3-dimensional Euclidean distance.

Figure 4 shows plots for the locations of cities, obtained from BMDS with dimension $p = 3$, and rotated manually to approximately fit the true location of the cities. One can observe that the cities are located approximately on the surface of a sphere with the radius of the earth and that the locations of the cities are well recovered.

5.3 Careers of Lloyds Bank Employees, 1905–1950

Sociologists are interested in characterizing and describing careers, to answer questions such as: What are the typical career patterns in a given period in a particular society? Have they been changing over time? Have people become more mobile occupationally?

One approach to doing this views a career as a sequence of occupations held, for example, in successive years, and then seeks to measure the similarity or dissimilarity between different

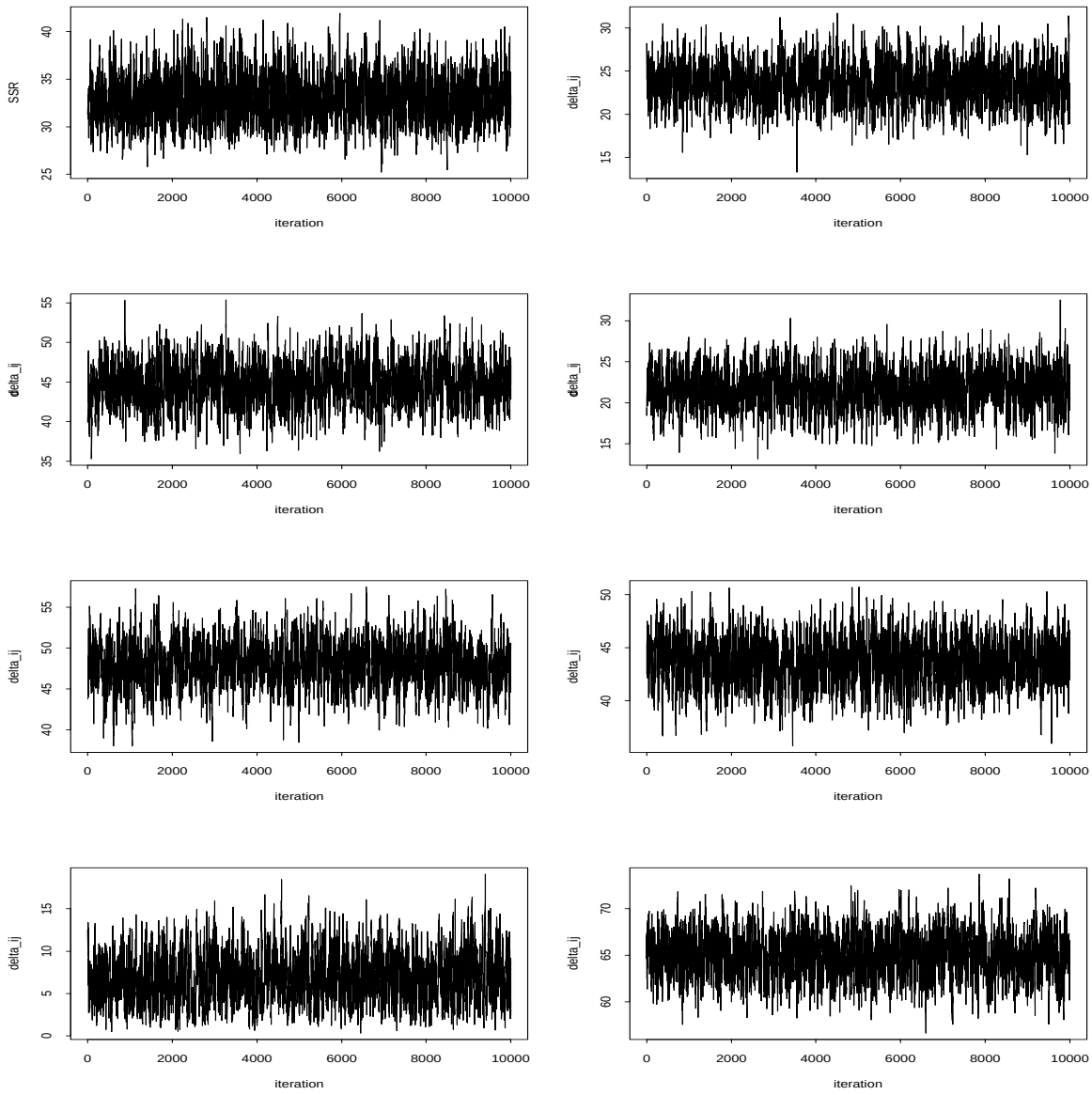


Figure 2: Time sequence plots of SSR and some δ_{ij}'s from MCMC iterations after burn-in for the City data.

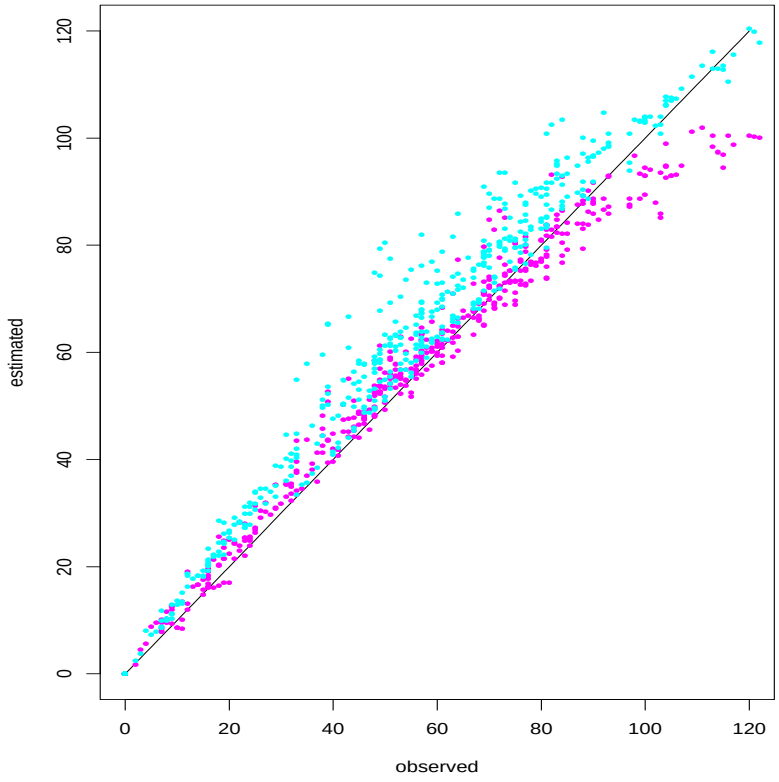
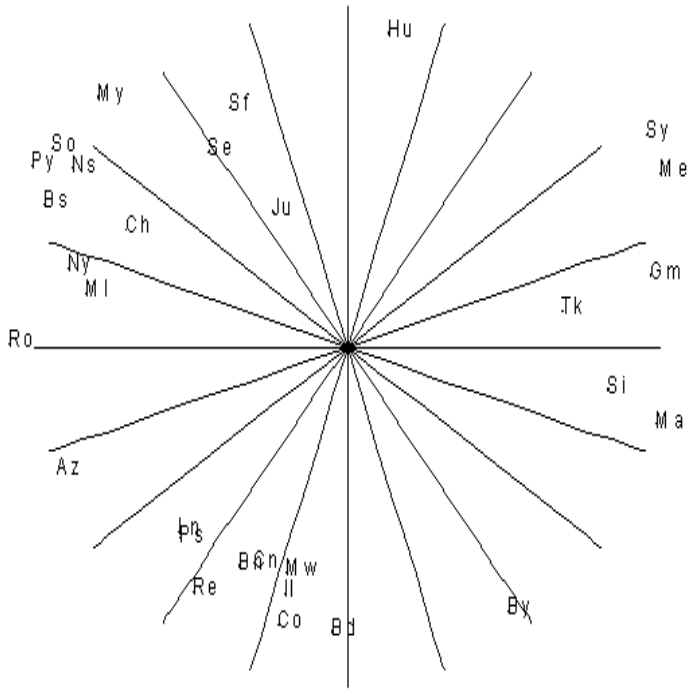
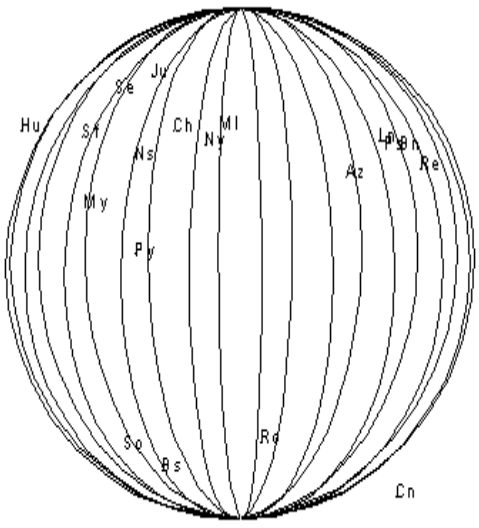


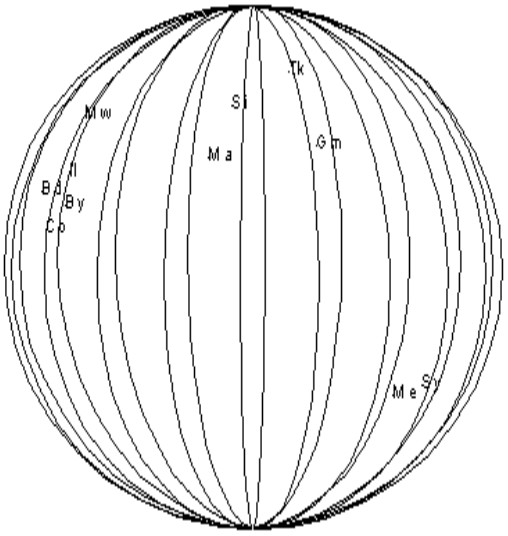
Figure 3: Observed and estimated distances for the Airline distance data (in units of 100 miles). The estimated distances from BMDS are represented by red dots and those from CMDS by green dots.



(a) View from the north pole.



(b) Hemisphere I



(c) Hemisphere II

Figure 4: Estimated locations of the cities from BMDS.

careers. Abbott and Hrycak (1990) proposed measuring the dissimilarity between the careers of two individuals by counting the minimum number of insertions, deletions and replacements that would be necessary to transform one career into another. Costs are associated with each kind of change, and the dissimilarity between the two careers is then measured as the total cost of transforming one career into another. This approach, known as optimal alignment, is borrowed from molecular microbiology, where it is applied to the comparison of DNA and protein sequences (Sankoff and Kruskal 1983; Boguski 1992).

Here we reanalyze some data considered by Stovel et al (1996), consisting of the careers of 80 randomly selected employees of Lloyds Bank in England, whose careers started between 1905 and 1909. This is part of a much larger study aimed at discovering how career patterns in large organizations have evolved over the course of the twentieth century. The more immediate goal here is to discover what the typical career sequences are, for data reduction and exploratory purposes, and also as a basis for further analysis. For each employee, information about his work position is available for each year he was at Lloyds. The information consists of the nature of the position (four categories, from clerk to senior manager), and the kind of place they were in (six categories, from small rural place to large city).

From the sequence data, an 80×80 matrix of dissimilarity measures was obtained, using the method of Abbott and Hrycak (1990); for more details, see Stovel et al (1996). Clearly the dissimilarities are not Euclidean distances, and may not satisfy certain geometric properties that hold for Euclidean distances, such as the triangle inequality. Our approach is to model the dissimilarities as before, with the idea that the non-Euclidean nature of the dissimilarities can be modeled at least approximately as part of the error. As we will see, this supposition turns out to be reasonable in practice.

We applied BMDS to the dissimilarity data. Table 3 presents the results of the analysis together with STRESS from CMDS. Again, BMDS performed much better than CMDS especially when the dimension is too small or too large. The improvement in performance of BMDS is more pronounced in this example than in the two previous examples. This suggests that BMDS is more robust than CMDS to variations in the alleged dimension and to violations of the Euclidean model assumption.

Dimension 8 is chosen as optimal since MDSIC attains its minimum at 8. Thus, the estimated configuration $\mathbf{X}$ when $p = 8$ can be used as a final estimate of $\mathbf{X}$. Figure 5 shows the fitted and observed dissimilarities for both BMDS and CMDS. The BMDS fitted dissimilarities fit the observed ones very well, considerably better than the CMDS ones; the sum of squared residuals for BMDS is less than half that for CMDS.

Figure 6 gives pairwise scatter plots of the first four dimensions of the BMDS estimates of $\mathbf{X}$.

dim	CMDS STRESS	BMDS STRESS	LRT	Penalty	MDSIC
1	0.5357	0.3545	-4228.1	325.8	26924
2	0.3390	0.1815	-3380.3	315.7	23022
3	0.2190	0.1063	-2924.9	310.9	19957
4	0.1280	0.0669	-1540.1	317.5	17343
5	0.0891	0.0524	-941.2	330.5	16120
6	0.0725	0.0452	-600.9	326.7	15510
7	0.0619	0.0411	-392.7	330.3	15236
8	0.0558	0.0386	-221.8	330.5	15173*
9	0.0547	0.0372	27.8	367.6	15282
10	0.0556	0.0374	7.7	396.1	15677
11	0.0600	0.0375	-17.3	310.3	16081
12	0.0637	0.0374	-21.2	444.5	16374

Table 3: Analysis of the LLoyd Bank data. The minimum MDSIC is marked with a star.

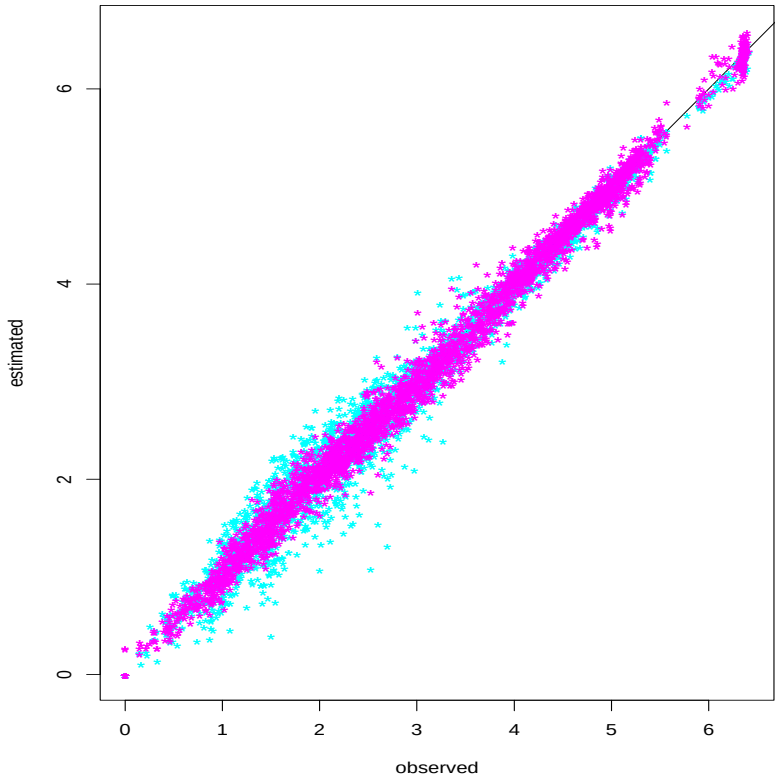


Figure 5: Fitted and observed dissimilarities for the LLoyd bank data. The red dots represent BMDS and the green dots correspond to CMDS.

The fourth dimension clearly separates two outliers. On closer inspection of the data it turned out that these were both individuals who had very short careers at Lloyds. They spent only a few years there, whereas all the other employees were at Lloyds for at least ten years.

The sociologists' interest in these data is primarily to characterize the typical career patterns at Lloyds in this period. To try to answer this question, we applied model-based clustering (Banfield and Raftery 1993; Fraley and Raftery 1998) to the BMDS estimate of $\mathbf{X}$, after removing the two clear outliers. This models the data as a mixture of multivariate normals, allowing for possible geometrically-motivated constraints on the covariance matrices of the different groups. The number of groups and the clustering model are both chosen using approximate Bayes factors, approximated via BIC.

Model-based clustering clearly identified three groups. These are shown in Figure 7, which displays the first two components of the BMDS solution. The three groups selected make clear substantive sense: Group 1 consists of 16 employees who had shorter careers (22 years or less), and spent all or almost all of their career at the lowest clerk rank. Group 2 consists of 30 employees with long careers (40 years or more), almost all of whom ended their careers at the lowest clerk level. Group 3 consists of 32 employees, most of whom were promoted and ended their careers as managers.

6 Summary and Discussion

In this paper, we have proposed a Bayesian approach to object configuration in multidimensional scaling and a simple Bayesian dimension criterion, MDSIC, based on the estimated object configuration. Bayesian MDS provided a better fit than classical MDS (Torgerson, 1952,1958) in all the cases we tried. The improvement in performance of BMDS is more pronounced when the dissimilarities are different from Euclidean distances and the effective dimension is ambiguous. This sort of robustness is useful in practice, since in applications dissimilarities are often not Euclidean distances and the concept of dimension may not even arise in their formulation. Another consideration is that one may often want to use two or three dimensions for visual display, although the true dimension may be much higher. The proposed dimension selection criterion, MDSIC, is easy to compute and gives a direct indication of optimal dimensionality. An advantage of MDSIC is that it uses the BMDS output, which seems to give good object configurations even when some of the model assumptions are violated.

A key feature of BMDS is that when the dimension increases, the coordinates for lower dimensions are changed, whereas in CMDS the coordinates for a lower dimension are always a subset of those for a higher one. The coordinates obtained from the lower dimensions are not necessarily

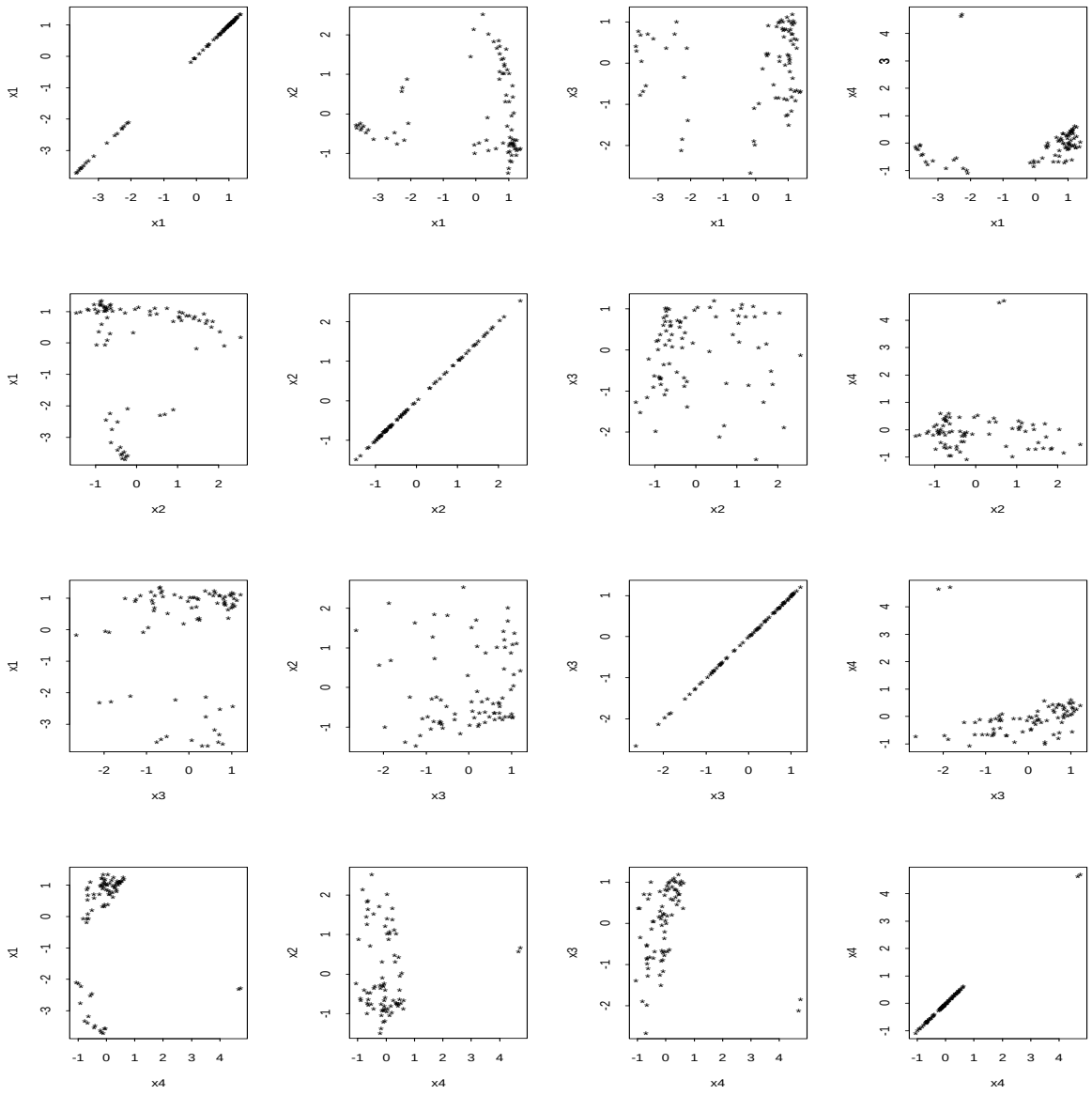


Figure 6: Pairwise scatter plots of the estimated object configuration from BMDS for the LLoyd Bank data.

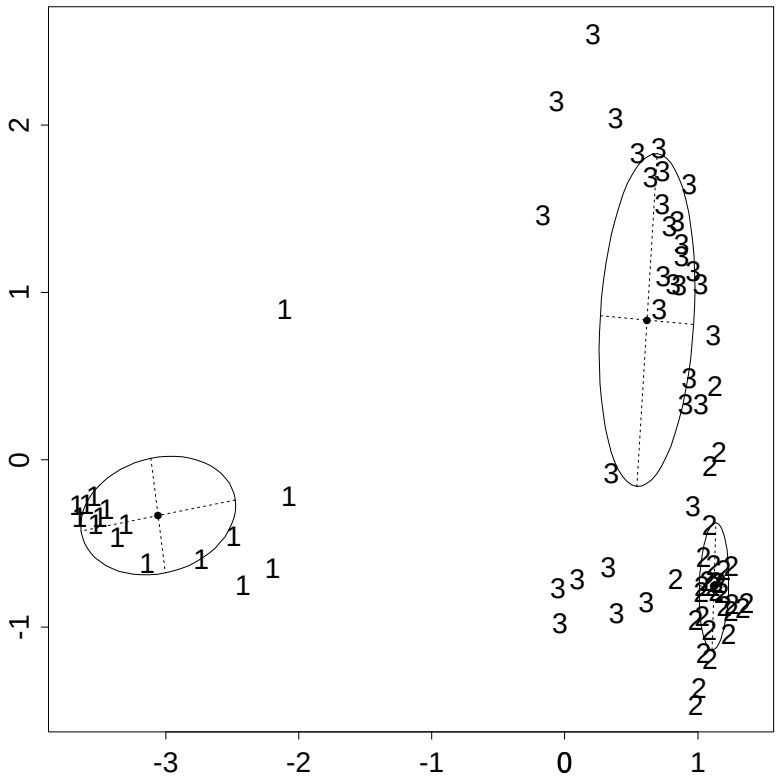


Figure 7: Lloyds bank Data: First two BMDS dimensions, with the 3-group model-based clustering classification shown. The ellipses show the one-standard-deviation contours of the densities of each of the component multivariate normal distributions, and the dotted lines show their principal axes.

an optimal choice when the dimension increases, and retaining them in higher dimensions may adversely affect the performance of CMDS.

One common reason for doing MDS is to cluster the objects. In our third example, we showed how model-based clustering can be used to do this, providing a formal basis for choosing the number of groups. The results were substantively reasonable and useful. Combining BMDS and model-based clustering thus provides a fully model-based approach to clustering dissimilarity data, including ways to choose the dimension of the data and the number of groups.

A more comprehensive approach to this problem would be to build a single model and carry out Bayesian inference for it. This could be done by using a prior distribution of $\mathbf{X}$ that is based on a mixture of multivariate normal distributions, rather than a single one as here. Then MCMC could be used to estimate both object configuration and group membership simultaneously. This approach could also provide a way of choosing the dimension and the number of groups simultaneously, rather than sequentially, as we did in our example. This seems desirable because there may be a tradeoff between dimension and number of groups. A maximum likelihood approach to the problem of clustering with multidimensional scaling of two-way dominance or profile data was proposed by DeSarbo et al (1991), but this is somewhat different from the present context, where the data come in the form of dissimilarities.

We have modeled dissimilarities as being equal to Euclidean distances plus error. This corresponds to metric scaling, and so our approach would perhaps best be called Bayesian *metric* multidimensional scaling. There has been a great emphasis in the MDS literature on nonmetric scaling, however. In nonmetric scaling, dissimilarities are modeled as equal to a nonlinear function of distance. This could be incorporated in the present framework by replacing (3) by

$$d_{ij} \sim N(g(\delta_{ij}), \sigma^2)\ I(d_{ij} > 0), \quad i \neq j, i, j = 1, ..., n, \tag{17}$$

where $g(\cdot)$ is a nonlinear but monotonic function. One could postulate a parametric model, or a family of parametric models, for g; one such family of models was proposed by Ramsey (1982). Then standard Bayesian inference via MCMC would again be possible, leading to Bayesian nonmetric multidimensional scaling.

Apart from the present work, we do not know of any other Bayesian analyses of multidimensional scaling for dissimilarity data. DeSarbo et al (1999) proposed a Bayesian approach to multidimensional scaling when the data are in the form of binary choice data, but this is rather different from the present context, where the data take the form of dissimilarities.

References

[1] Abbott, A. and Hrycak, A. (1990), "Measuring Sequence Resemblance," *American Journal of Sociology*, 96, 144–185.

[2] Banfield, J.D., and Raftery, A.E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803-821.

[3] Boguski, M.S. (1992), "Computational Sequence Analysis Revisited," *Journal of Lipid Research*, 33, 957–974.

[4] Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*, Springer-Verlag, New York, Berlin.

[5] Cox, D.R. (1982), Comment, *Journal of the Royal Statistical Society, Series A*, 145, 308–309.

[6] Cox, T.F. and Cox, M.A.A. (1994). *Multidimensional Scaling*, Chapman & Hall, London.

[7] Davison, M.L. (1983), *Multidimensional Scaling*, Wiley, New York.

[8] DeSarbo, W.S., Howard, D.J., and Jedidi, K. (1991), "MULTICLUS — A New Method for Simultaneously Performing Multidimensional Scaling and Cluster Analysis," *Psychometrika*, 56, 121–136.

[9] DeSarbo, W.S., Kim, Y., and Fong, D. (1999), "A Bayesian Multidimensional Scaling Procedure for the Spatial Analysis of Revealed Choice Data," *Journal of Econometrics*, **89**, 79-108.

[10] Fraley and Raftery (1998), "How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis," *Computer Journal*, **41**, 578-588.

[11] Gelman A., Roberts, G.O., and Gilks, W.R. (1996), "Efficient Metropolis jumping rules," *Bayesian Statistics*, **5**, 599-608.

[12] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

[13] Groenen (1993), *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*, Lieden, The Netherlands: DSWO Press.

[14] Groenen, Mathar, and Heisser, W.J. (1995), "The Majorization Approach to Multidimensional Scaling for Minkowski Distances," *Journal of Classification*, **12**, 3-19.

[15] Kass, R.E. and Wasserman, L.A. (1995), "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

[16] Kruskal, J.B. (1964), "Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis," *Psychometrika*, 29, 1–28.

[17] Hastings, W.K. (1970), "Monte Carlo Sampling Methods using Markov Chains and their Applications," *Biometrika*, **57**, 97-109.

[18] MacKay D. (1989), "Probabilistic Multidimensional Scaling: An Anisotropic Model for Distance Judgements," *Journal of Mathematical Psychology*, **33**, 187-205.

[19] MacKay D. and Zinnes, J.L. (1986), "A Probabilistic Model for the Multidimensional Scaling of Proximity and Preference Data," *Marketing Sciences*, **5**, 325-334.

[20] Raftery, A.E. (1995), "Bayesian Model Selection in Social Research (with Discussion)," *Sociological Methodology*, 25, 111–193.

[21] Raftery, A.E. (1999), "Bayes Factors and BIC - Comment on 'A critique of the Bayesian information criterion for model selection'," *Sociological Methods and Research*, 27, 411–427.

[22] Ramsay, J.O. (1982), "Some Statistical Approaches to Multidimensional Scaling," *Journal of the Royal Statistical Society A*, **145**, 285-312.

[23] Roberts, G. O., Gelman, A., and Gilks, W. R. (1997), "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms," *Annals of Applied Probability*, **7**, 110–120.

[24] Sankoff, D., and Kruskal, J.B. (1983), *Time Warps, String Edits, and Macromolecules*, Reading, Mass.: Addison-Wesley.

[25] Schutze, H. and C. Silverstein (1997), "Projections for efficient document clustering", *ACM SIGIR 97*, pp. 74-81.

[26] Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–466.

[27] Stovel, K., Savage, M., and Bearman, P. (1996), "Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970," *American Journal of Sociology*, 102, 358–399.

[28] Takane, Y.(1982), "The Method of Triadic Combinations: A New Treatment and Its Applications," *Behaviormetrika*, **11**, 37-48.

[29] Takane and Carroll, J.D. (1981), "Nonmetric Maximum Likelihood Multidimensional Scaling from Directional Rankings of Similarities," *Psychometrika*, **46**, 389-405.

[30] The World Almanac (1966).

[31] Tibshirani, R., Lazzeroni, L., Hastie, T., Olshen, A., and Cox, D. (1999), "The Global Pairwise Approach to Radiation Hybrid Mapping," Technical Report, Department of Statistics, Stanford University.

[32] Torgerson, W.S. (1952), "Multidimensional Scaling: I. Theory and Method," *Psychometrika*, **17**, 401-419.

[33] Torgerson, W.S. (1958), *Theory and Methods of Scaling*, Wiley, New York.

[34] Young, F.W. (1987), *Multidimensional Scaling - History, Theory, and Applications*, Erlbaum Association, Hillsdale.

APPENDIX

Justification of (11)

Integration of $h(\sigma^2, \mathbf{X})$ where

$$h(\sigma^2, \mathbf{X}) = (\sigma^2)^{-(m/2+a+1)} \exp\left[-\frac{SSR/2+b}{\sigma^2} - \sum_{i>j} \log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]$$

is not straightforward. However, in most cases $m = n(n-1)/2$ is very large and the likelihood of σ^2 dominates the prior and hence $h(\sigma^2, \mathbf{X})$ is approximately proportional to the likelihood

$$l(\sigma^2, \mathbf{X}) \equiv (\sigma^2)^{-m/2} \exp\left[-\frac{SSR}{2\sigma^2} - \sum_{i>j} \log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]. \tag{18}$$

In addition, because of the large m, the likelihood $l(\sigma^2, \mathbf{X})$ is well approximated by a normal density function. Thus, applying a Laplace approximation to the integral of $l(\sigma^2, \mathbf{X})$ gives

$$\int h(\sigma^2, \mathbf{X}) d\sigma^2 \approx \int l(\sigma^2, \mathbf{X}) d\sigma^2 \approx (2\pi)^{1/2} H^{-1/2} l(\mathbf{X}, \hat{\sigma}^2), \tag{19}$$

where H is the minus Hessian of the log likelihood and $\hat{\sigma}^2$ is the MLE of σ^2.

We now argue that the probability $\Phi(\delta_{ij}/\sigma)$ is unlikely to have much effect on the model comparison, and can safely be ignored. Suppose we are comparing dimension p with dimension $(p+1)$. We distinguish between two situations. Suppose first that the true dimension is $(p+1)$.

Then, asymptotically, the term $(-\text{SSR}/2\sigma^2)$ will dominate the exponent on the right-hand side of (18), dimension $(p+1)$ will be preferred, and the term $\sum_{i>j} \log \Phi(\delta_{ij}/\sigma)$ will be immaterial. Second, suppose instead that the true dimension is p. Then the fitted δ_{ij} will be the same, asymptotically, for dimension p as for dimension $(p+1)$, and so the term $\sum_{i>j} \log \Phi(\delta_{ij}/\sigma)$ will be the same for both dimensions. Thus it will cancel in the comparison, and can again be ignored.

Thus, we ignore the term $\sum_{i>j} \log \Phi(\delta_{ij}/\sigma)$ and use the approximation

$$l(\sigma^2, \mathbf{X}) \approx l^*(\sigma^2, \mathbf{X}) \equiv \ (\sigma^2)^{-m/2} \ \exp\left[-\frac{SSR}{2\sigma^2}\right].$$

Replacing l by l^* and H by the minus Hessian H^* of l^* in (19) and letting $\hat{\sigma}^2 = SSR/m$ which maximizes l^* gives the formula (11).